Counterfactual Sepsis Outcome Prediction Under Dynamic and Time-Varying Treatment Regimes

Megan Su[1]*, Stephanie Hu[1]*, Hong Xiong[2]*, Elias Baedorf Kassis[3], Li-wei H Lehman[1]†
[1] Massachusetts Institute of Technology, Cambridge, MA; [2]Harvard University, Cambridge, MA; [3] Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

Abstract

Sepsis is a life-threatening condition that occurs when the body's normal response to an infection is out of balance. A key part of managing sepsis involves the administration of intravenous fluids and vasopressors. In this work, we explore the application of G-Net, a deep sequential modeling framework for g-computation, to predict outcomes under counterfactual fluid treatment strategies in a real-world cohort of sepsis patients. Utilizing observational data collected from the intensive care unit (ICU), we evaluate the performance of multiple deep learning implementations of G-Net and compare their predictive performance with linear models in forecasting patient outcomes and trajectories over time under the observational treatment regime. We then demonstrate that G-Net can generate counterfactual prediction of covariate trajectories that align with clinical expectations across various fluid limiting regimes. Our study demonstrates the potential clinical utility of G-Net in predicting counterfactual treatment outcomes, aiding clinicians in informed decision-making for sepsis patients in the ICU.

1 Introduction

Under real-world conditions, we can only observe one set of outcomes: that is, outcomes which occurred under the observational regime. However, often we might wonder what might have happened had a different course of action been followed. This is particularly important for clinicians who may have to choose between multiple treatment options for their patients but do not have the ability to test every strategy before making a decision.

Sepsis patients, who frequently display heterogeneous responses to similar therapies, may particularly benefit from clinicians being able to predict the effects of different intervention strategies prior to administration. This task is known as *counterfactual prediction*, in which the goal is to estimate the trajectories of potential outcomes under different interventions given previous observed covariate history [1, 2].

In this study, we apply G-Net [3], a deep learning approach to g-computation, to predict the outcomes of alternative fluid resuscitation treatment regimes on sepsis patients in the intensive care unit (ICU). G-computation [1] is a causal inference method for estimating expected counterfactual outcomes under dynamic, time-varying treatment strategies. Our work builds on [3] by exploring different implementations to G-Net and adapting the model to make predictions on real-world observational data from the MIMIC database [4].

We investigate G-Net's predictive performance when varying-length of a patient's history was used as input to the model in predicting future outcomes under observational treatment regimes. We compare the performance of deep learning models with traditional linear models, the latter of which has conventionally been utilized in g-computation [5, 6]. We then employ G-Net to forecast treatment responses in sepsis patients under different fluid-limiting strategies, incorporating varied caps on total volume of fluid boluses administered during the initial 24 hours in the ICU. Our fluids-limiting strategies were inspired by [7] and the recently conducted randomized controlled trial, Crystalloid Liberal or Vasopressors Early Resuscitation in Sepsis (CLOVERS) [8], a randomized clinical trial studying outcomes of restrictive vs. liberal fluid management strategies in sepsis patients. Finally, we present an individual case study from the MIMIC-IV database [4] to illustrate the potential use of G-Net for personalized treatment response prediction at the individual patient level.

2 Related Work

Due to their ability to handle complex time dependencies between variables, there has been great interest in applying recurrent neural networks (RNNs) to the problem of estimating time-varying treatment effects. Lim et al. explored the

[1]* indicates co-first authors. † Corresponding author: lilehman@mit.edu.

use of RNNs in implementing marginal structural models (so-called recurrent MSMs) for counterfactual predictions and demonstrated that their model outperformed linear baselines and traditional MSM approaches [9]. Recent research have also looked at predicting outcomes under counterfactual strategies given data under the observational regime [10, 11, 12, 13]. Bica et al. [13] introduced the Counterfactual Recurrent Network (CRN) which applies domain adversarial training to build treatment-invariant representations for estimating effects of counterfactual interventions.

Although these techniques show promise in estimating treatment effects, they are restricted to either learning point exposures and/or making outcome predictions for time-varying treatment strategies that are static, where treatment decisions do not depend on covariate history. In contrast, we are interested in predictions for *dynamic* regimes, where treatment decisions depend on time-varying covariates. G-computation relies on different modeling assumptions than MSMs, and can handle high dimensional health history in particular. It is also able to estimate the distribution of counterfactual outcomes under a time-varying treatment strategy, which is not as straightforward to do with MSMs. The g-computation algorithm [2] learns observational covariate distributions conditioned on past history. While previous studies have mostly relied on generalized linear models (GLMs) for this task [5, 6], there is no conceptual barrier to substituting these models with more complex ones. Prior works have investigated implementations of counterfactual predictions using Gaussian processes [14, 15] and Transformers [16], but did not focus on applications to dynamic and time-varying treatment settings using real-world data. Recent work by [17] presented an alternating sequential model that used Transformers for clinical outcome prediction, but their work did not support counterfactual prediction.

Research by [3] developed G-Net, a flexible sequential deep learning framework for g-computation, to estimate the conditional distribution of covariates given patient history under dynamic, time-varying treatment strategies; however, [3] only evaluated G-Net on synthetic datasets. Our study extends upon our prior work by modifying G-Net for counterfactual sepsis treatment outcome prediction using a real-world dataset. Specifically, we experiment with a different architectural design and test different model implementations to simulate the distributions of multivariate covariates.

3 Methods

We adapted the G-Net framework introduced by [3] to model real-world data from the MIMIC database [4]. We train and evaluate our models to assess their predictive performance under the observational treatment regime. We then applied the learned models in predicting patient trajectories and outcomes under different counterfactual treatment strategies in sepsis patients.

Dataset The data employed in this work was extracted from the Medical Information Mart for Intensive Care IV database (MIMIC-IV v1.0), containing medical records from more than 523,500 hospital admissions and 76,500 ICU stays at the Beth Israel Deaconess Medical Center (BIDMC) between 2008 and 2019 [4]. Our cohort was limited to ICU stays in which the patient was identified as septic under the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) [18]. Patients with missing records of pre-ICU fluids or admitted to the ICU following cardiac, vascular, or trauma surgery were removed. The final dataset consisted of 8,721 ICU stays, each associated with a distinct patient.

As predictors to our model, we included 45 physiological and clinical time-varying covariates that are typically monitored in the ICU and important for determining sepsis intervention strategies, including potential confounders. Patient demographics, co-morbidities and pre-ICU fluid bolus volumes were included as static variables in our model. Our treatment variable was fluid bolus volume, where we define a *bolus* as fluid administered at a rate of 250mL/hr or greater [4]. To avoid possible confounding, all other fluids documented for a patient were treated as "maintenance fluids". The outcomes of interest in our study were fluid overload, which was assessed by identifying events commonly associated with the condition, including administration of diuretics, onset of dialysis, initiation of mechanical ventilation, and chest X-ray findings of pulmonary edema.

Model In our study, we aimed to estimate outcomes under dynamic, time-varying counterfactual treatment strategies given observed patient histories with high-dimensional histories. *Time-varying* describes treatments that comprise decisions at multiple time points while *dynamic* indicates that the intervention at each time point is dependent on the history up to that time point. G-computation [2] is an algorithm that is particularly well-suited to this task.

Given data under the observational regime, g-computation (1) estimates the conditional distribution of relevant covariates given covariate and treatment history at each time point, then (2) produces Monte Carlo estimates of counterfactual outcomes by simulating forward patient trajectories under treatment strategies of interest [2, 1]. The G-Net framework [3] enables the use of sequential deep learning models to estimate conditional covariate distributions in step (1), which can then be used to predict covariate trajectories in step (2). We refer readers to [3] for a detailed description of the G-Net framework and the assumptions required for g-computation.

In this study, we implemented a G-Net architecture in which each time-varying covariate is trained and optimized separately using an individual deep sequential model when learning covariate distribution. More generally, when learning covariate distributions, one can separately model multivariate covariates into p disjoint groups or components, where each group is just a potentially multivariate subset of the covariate vector that is jointly modeled in the same model in our architecture. When $p = 1$, one models all covariates simultaneously and directly approximate the joint covariate distribution. On the other hand, by setting $p > 1$, we can model groups of covariates separately. This may be desirable if covariates have different distribution types, as in the case of mixing continuous and categorical variables, rendering it difficult to sample from their joint distribution. In [3], we implemented G-Net where $p = 2$, with one deep model for continuous variables, and another one to model the categorical variables. In training separate models for different covariates, we aim to improve performance in modeling conditional distribution for covariates.

We compare 4 variants of G-Net: one in which all variables are modeled using GLMs, basic recurrent neural network (RNN), RNNs using gated recurrent units (GRU), and long short-term memory (LSTM) networks. We tested a variety of recurrent architectures to emulate conditioning on patient history, since a feature of our chosen recurrent architectures is their use of a hidden state to store memory. The GLM model was used as a baseline for comparison.

Model Training We divided our sepsis cohort into training, validation, and testing sets using an 80-10-10 split, respectively, and trained each of the four variants of G-Net on a one-step-ahead prediction task. The model with the highest validation performance was used for later experiments. The goal of training was to approximate the distribution for each covariate so that we could use the model to simulate forward those covariate distributions during testing.

Clinical Outcome Prediction Oftentimes, the effects of treatment are not immediate and only appear after 2 or 3 days after administration. Because of this, we elected to assess clinical outcomes occurring within 72 hours of ICU admission for patients who did not experience outcomes in the first 24 hours. For each outcome of interest, we implemented a binary classifier, modeled using LSTMs, to predict whether the patient would experience that outcome within 72 hours of ICU admission, conditioned on the patient's 24-hour covariate trajectories. There were six outcomes of interest: diagnosis of pulmonary edema, on mechanical ventilator, dialysis, administration of diuretics, ICU release, and in-hospital mortality. We impose an arbitrary ordering for predicting the 72-hour outcomes such that we also condition on the predictions for outcomes 0 to $j - 1$ when predicting outcome j. However the output of our classifier will be a continuous value in the range $[0, 1]$. We thus threshold our predictions to obtain binary values to serve as input to the next classifier. Specifically, we use a threshold found via the validation dataset by picking the optimal threshold on the validation ROC curve.

The same training, validation, and testing splits that were used to train the different G-Net models were used to train the 72-hour outcome classifiers, with the exception of patients who already experienced the outcome of interest in the first 24 hours. The outcome ground-truth labels were extracted by looking for documentation of the outcome in each patient's 72-hour ICU data. The ground truth 24-hour covariate trajectories from the training set were used as inputs to train our 72-hour outcome prediction models. During testing, we input the simulated trajectories from 100 Monte Carlo simulations for each test set patient to generate a predicted probability for the target outcome of interest.

Experiments We first evaluated our models' performance via *predictive check*, in which we performed Monte Carlo simulations and simulated forward patient covariate trajectories in the test set under the observational regime and compared the simulated trajectories (averaged across 100 Monte Carlo simulations per patient) to ground-truth data. During testing, we experimented with different simulation start times post-ICU admission, denoted k. When simulating with a start time k, we fed the first $k - 1$ timesteps of a test patient's ground-truth trajectory (including treatments) into G-Net before initiating the feed-forward simulation process at timestep k. This represents the scenario in which a clinician who has been observing a patient for some time in the ICU would like to predict what will happen next, conditioned on the patient's first $k-1$ timesteps in the ICU.

To predict clinical outcomes within 72 hours of ICU admission, we fed the predicted 24-hour covariate trajectories as input to individual outcome classification models. We interpreted the output of each classifier as the probability that the patient would experience the outcome within 72 hours of ICU admission. The final predicted probability of a patient experiencing the outcome was the average of the predicted probabilities across all Monte Carlo-simulated trajectories for that patient.

Following the predictive check, we assessed our model in performing counterfactual prediction. The feed-forward simulation strategy was similar to that used in the predictive check. However, the treatments at each time step are administered based on the treatment rules as defined by the counterfactual regime, conditioned on the patient covariate and treatment history.

Predictive Check Evaluation We use root mean square error (RMSE) to quantify differences between predicted and observed trajectories across time intervals (up to 24 hours) of individual patients. For each patient, the predicted trajectory is the average of 100 Monte Carlo simulations from G-Net conditioned on the observed history of the patient up until time (k-1).

To accommodate the variability in patient trajectory lengths due to mortality or discharge, and to address potential over- or under-predictions in trajectory length compared to ground truth, we propose the following approach for calculating individual-level RMSE. First, for patients who died within the first 24 hours, we replace all subsequent time steps of continuous variables with zero (normalized) until the 24th hour. Covariates with potential negative values are filled with the normalized minimum value in the dataset, while covariates following a log-normal distribution are filled with a normalized logarithmic zero. Second, for patients who were released from the hospital before the end of the first 24 hours in the ICUs, we fill all time steps following hospital release with the population-mean for all continuous variables (up until hour 24). Third, we do the same for the predicted chains (after the predicted death and release outcomes during the first 24 hours) so that all chains have lengths of 24-hours. We then compute the average RMSE across all patients by comparing the predicted trajectory (averaged across 100 Monte Carlo simulations per patient) and the ground-truth trajectory of each patient across all continuous variables.

Following comparisons of 24-hour trajectories between predicted and ground-truth, we compared 24-hour and 72-hour outcomes. We primarily focused on area under the receiver operating characteristic (ROC) curve (AUC) as a measure of outcome prediction performance. Confidence intervals around AUCs are generated based on techniques described in [19]. To generate a probability for the 24-hour window, we first derived a binary label for each Monte Carlo simulation based on whether the outcome is predicted to have occurred: 1 if outcome of interest was predicted to occur between hour k to 24, and 0 otherwise. We then used the average across the Monte Carlo simulations per patient as the predicted probability for that patient. We determined the ground truth in the same manner by looking for the presence of the outcome of interest in the ground-truth trajectory.

To generate the final predicted 72 hour probability for a patient, for each Monte Carlo simulation, we label a trajectory with probability 1 if the patient experienced the outcome in the first 24 hours, probability 0 if the patient ended their stay in the first 24 hours without experiencing the outcome, or otherwise feed the trajectory into our binary classifier to find a predicted probability. Then we take the average across the simulations to get a final predicted probability, which was then compared against the ground-truth label of if the patient experienced the outcome in the first 72 hours. For both the 24 hour and 72 hour evaluations, we excluded a patient from the calculations if the patient had ended their stay from time steps 0 to $k-1$ because we would have already seen the final outcomes that the patient experienced. For the 72-hour evaluations, we additionally excluded patients who any of the outcomes of interest during timestep 1 to $k-1$, who would be out of distribution from our training set.

Counterfactual Experiments We adapted our counterfactual (CF) strategies from established clinical trials studying the early treatment of sepsis. More specifically, our CF strategies were inspired by the Crystalloid Liberal or Vasopressors Early Resuscitation in Sepsis (CLOVERS) clinical trial [8]. In our CF strategy, for a patient with blood pressure below 65mmHg at time t, a 1L bolus was administered if the total volume of fluids (including both treatment and maintenance) they received up until that time point did not exceed X liters and if they did not exhibit any signs of fluid overload (as indicated by the presence of pulmonary edema based on chest x-ray radiology reports). Our fluids con-

Characteristic	***n***
Number of ICU stays	8,721
Age, mean (std)	65.31 (17.33)
Pre-ICU Fluids, mean (std)	2.87 L (1.99L)
Gender (% Male)	56.30%
Race – white	66.40%
Race – black	8.78%
Race – other	24.81%

Table 1: Statistics describing our entire Sepsis-3 cohort, including age, gender, and race. .

Outcome	**24 hr**	**72 hr/hosp**
Pulmonary Edema (%)	32.85	44.36
Mechanical Ventilation (%)	42.46	45.08
Diuretics (%)	15.46	27.19
Dialysis (%)	2.56	4.55
Release (%)	11.49	86.00*
In-Hospital Mortality (%)	1.83	14.00*

Table 2: Proportion of population that experiences each outcome within 24 hours of ICU admission and 72 hours of ICU admission respectively. * Release and mortality refer to hospital discharge and mortality respectively.

servative CF strategies focus on fluids bolus administration. Maintenance fluids were withheld for fluids-conservative strategies. For all CF strategies, vasopressor administration was modeled as a confounding variable, and not explicitly controlled as part of our CF strategies. In our investigations, we experimented with imposing a fluid cap of 3L and 5L as fluid conservative strategies. We also experiment with a fluid CF strategy with no fluid cap to represent a fluid liberal strategy.

Individual case study We chose to investigate individuals in our test set who suffered from a history of congestive heart failure (CHF), a condition known to present challenges in sepsis fluids administration, as over-administration of fluids can worsen their condition, and lead to adverse outcomes such as pulmonary edema. We modelled the effects of a fluid liberal counterfactual strategy compared to a fluid conservative strategy.

4 Results

Cohort Statistics Our sepsis cohort consisted of 8,721 total patients which we further split into 6,972 patients, 872 patients, and 873 patients for the training, validation, and test data respectively. To get a better understanding of our cohort of interest, Table 1 presents our cohort summary statistics, including the demographics, Additionally, due to our interest in fluid strategies, we also present statistics on our cohort's pre-ICU fluids. Table 2 presents the proportion of the population who experienced outcomes of interest in the first 24 hours and first 72 hours post-ICU admission. We note that the hospital mortality and release (in the 72-hr/hosp column) refers to death and release in the hospital, and not restricting to the first 72-hours in the ICUs. We notice that the proportions of the population are similar to those of the test set, which will be presented later in the paper, indicating that our test set is fairly representative of the population.

Predictive Checks In the predictive check, we simulated forward covariate trajectories up to 24 timesteps under the observational regime, conditioned on the past $k - 1$ timesteps of treatment post-ICU admission, and compared the predicted trajectories against ground-truth trajectories in Table 3. We then fed the predicted trajectories into separate outcome classifier models to predict in-hospital mortality and other outcomes related to fluid overload occurring in the first 72 hours post-ICU admission. We evaluated our models' performance in predicting outcomes by comparing the predicted outcomes against the ground-truth outcomes under the observational regime. Finally, we used our models to simulate counterfactual trajectories, which we fed into the outcome classifiers. Due to lack of ground-truth outcomes under counterfactual regimes, we performed a qualitative analysis of the counterfactual predictions.

Predictive Check: Individual-Level Trajectories We see that the RMSE in predicted 24-hour patient trajectories is relatively stable across different values of k (Table 3) while predicted 24-hour clinical outcomes of interest tend to be more accurate as the value of k increases (Table 4). G-Net models implemented using RNN architectures (LSTM, GRU, and basic RNN units) generally outperform those implemented using GLMs. The LSTM implementation had the best overall performance.

Table 3: Predictive check on continuous variables. Performance of various G-Net architectures in predicting 24-hour patient covariate trajectories, starting at time k conditioned on covariates from the previous $k-1$ timesteps of post-ICU admission. k ranges from 2 to 12 hours after ICU admission. Values reported represent the RMSE of continuous covariates between the predicted and ground-truth trajectories. To account for mismatch in trajectory length due to over- and under-prediction of death and release time, for both predicted and actual trajectories, depending on nature of covariates, post-death timesteps are padded with one of the following values: normalized zero, normalized minimum value in dataset, or normalized logarithmic zero. The post-release timesteps are consistently padded with normalized population mean.

Model	**k = 2**	**k = 4**	**k = 6**	**k = 8**	**k = 10**	**k = 12**
Linear	5.949	6.232	6.792	6.570	6.584	7.084
GRU	**5.663**	6.226	6.211	6.373	6.699	**6.447**
RNN	5.842	**6.226**	6.053	6.223	6.490	6.800
LSTM	5.864	6.286	**5.921**	**5.765**	**6.439**	6.456

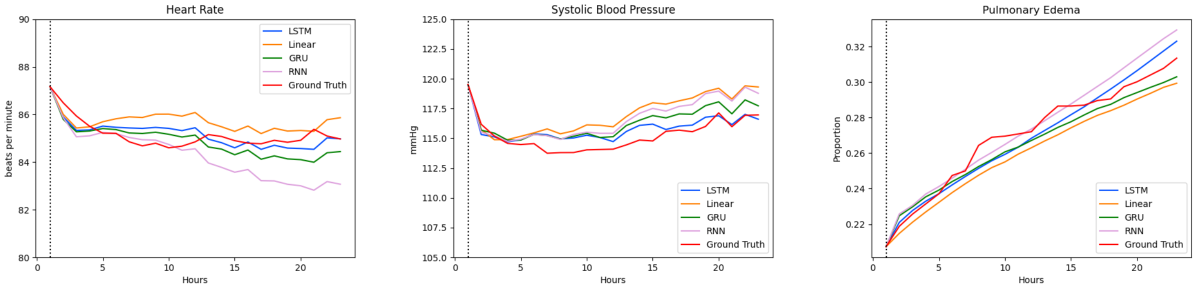

Figure 1: Comparison of simulated and ground-truth population-level trajectories for selected covariates in the predictive check experiments under different versions of G-Net.

Predictive Check: Population-Level Trajectories We present trajectories of selected covariates simulated with $k = 2$ in Figure 1. For each covariate, we show the population average of the predicted trajectories from the test set, with 100 Monte Carlo simulations per patient. Note that the number of trajectories contributing to the average could decrease over time due to death or discharge in the simulated trajectories. We found that the population-level predictions made by the different G-Net models aligned relatively closely with the ground-truth trajectories. We see

Table 4: Predictive Check. Performance of G-Net in predicting clinical outcomes within the first 24 hours of ICU admission at varying simulation start times. Values reported are AUCs and 95% confidence intervals. k = hours since ICU admission.

Model	k	**Pulmonary Edema**	**MV**	**Diuretics**	**Dialysis**
Cohort Statistics		(873, 30.24%)	(873, 38.72%)	(873, 15.35%)	(873, 1.95%)
Linear	2	0.85 (0.82, 0.88)	0.89 (0.87, 0.92)	0.74 (0.69, 0.79)	0.78 (0.67, 0.90)
GRU	2	**0.86** (0.83, 0.89)	**0.91** (0.88, 0.93)	0.77 (0.73, 0.82)	0.88 (0.79, 0.98)
RNN	2	0.85 (0.82, 0.88)	**0.91** (0.89, 0.93)	0.75 (0.70, 0.8)	**0.93** (0.86, 1.00)
LSTM	2	**0.86** (0.84, 0.89)	0.90 (0.88, 0.92)	**0.82** (0.77, 0.86)	0.92 (0.84, 1.00)
Linear	6	0.89 (0.86, 0.91)	0.93 (0.91, 0.95)	0.72 (0.66, 0.77)	0.87 (0.78, 0.97)
GRU	6	**0.90** (0.87, 0.92)	0.94 (0.92, 0.96)	**0.82** (0.78, 0.87)	0.92 (0.85, 1.00)
RNN	6	0.88 (0.86, 0.91)	0.94 (0.92, 0.96)	0.78 (0.74, 0.83)	**0.96** (0.90, 1.00)
LSTM	6	0.89 (0.87, 0.92)	**0.95** (0.94, 0.97)	0.81 (0.76, 0.86)	0.95 (0.89, 1.00)
Linear	8	**0.91** (0.89, 0.94)	0.93 (0.91, 0.95)	0.73 (0.68, 0.79)	0.97 (0.91, 1.00)
GRU	8	0.89 (0.87, 0.92)	0.95 (0.93, 0.97)	0.79 (0.74, 0.84)	0.92 (0.83, 1.00)
RNN	8	**0.91** (0.88, 0.93)	0.95 (0.93, 0.97)	0.78 (0.72, 0.83)	**0.98** (0.93, 1.00)
LSTM	8	**0.91** (0.88, 0.93)	**0.96** (0.94, 0.97)	**0.80** (0.75, 0.85)	0.90 (0.82, 0.99)

that as time progresses, the predicted population-level trajectories tend to deviate further from the ground truth.This is expected and can be attributed to compounding error for longer sequences.

Predictive Check: 72-Hour Outcomes Results in Table 5 show AUCs (and 95% CI) from using an LSTM to classify target clinical outcome of interest using the predicted 24-hour simulation trajectories generated by various models of G-Net (including Linear, GRU, basic RNN, and LSTM) under observational treatment regime. We find that G-Net models implemented using RNN-based architectures, especially LSTMs, tend to perform better than using linear models, particularly in predicting outcomes related to fluid overload within 72 hours of ICU stays.

Table 5: Performance of G-Net in predicting clinical outcomes within the first 72 hours of ICU admission at varying simulation start times. Values reported are AUCs and 95% confidence intervals.

Model	k	Pulmonary Edema	MV	Diuretics	Dialysis
Cohort Statistics		(873, 43.64%)	(873, 44.90%)	(873, 27.38%)	(873, 4.70%)
Linear	6	0.67 (0.62, 0.72)	**0.71** (0.64, 0.77)	0.67 (0.62, 0.72)	0.87 (0.80, 0.95)
GRU	6	**0.68** (0.63, 0.73)	0.67 (0.60, 0.74)	0.71 (0.66, 0.75)	0.89 (0.82, 0.96)
RNN	6	0.65 (0.60, 0.70)	**0.71** (0.64, 0.78)	0.70 (0.65, 0.74)	0.90 (0.83, 0.97)
LSTM	6	0.67 (0.62, 0.72)	**0.71** (0.64, 0.77)	**0.72** (0.68, 0.76)	**0.91** (0.84, 0.98)
Linear	8	**0.69** (0.64, 0.74)	0.65 (0.58, 0.72)	0.67 (0.62, 0.72)	0.90 (0.83, 0.97)
GRU	8	0.67 (0.62, 0.72)	0.72 (0.65, 0.79)	0.71 (0.66, 0.76)	0.91 (0.84, 0.98)
RNN	8	0.67 (0.62, 0.72)	0.72 (0.65, 0.79)	0.68 (0.63, 0.73)	0.90 (0.83, 0.98)
LSTM	8	**0.69** (0.64, 0.74)	**0.73** (0.66, 0.8)	**0.72** (0.67, 0.77)	**0.92** (0.86, 0.99)

Counterfactual Population Trajectories We used the LSTM version of G-Net to simulate covariate trajectories for patients in the testing set under two conservative fluids strategies (with 3L and 5L fluid cap) and a fluid liberal strategy (with no cap). Select covariates are presented in Figure 2, allowing for comparison of the three counterfactual regimes. Population-level average trajectories for the test set shown (N=873). Counterfactual strategies are applied starting at time 1, conditioned on observations up until the first hour in the ICU. For the population-level trajectory plots, we perform 5 Monte Carlo simulations per patient. Values plotted are averaged across simulated trajectories for all test patients. Our aim is to assess whether G-Net can make counterfactual predictions that are aligning with both physiological and clinical expectations. The trends depicted in Figure 2 align with clinical expectations: under fluid liberal strategies, the bolus volume, urine output, and blood pressure exhibit higher values, following the expected trends. This observation is in agreement with our expectations at a population level. Additionally, the lower lactate levels observed under the fluid liberal regime are in line with what one would predict. Conversely, the higher creatinine levels noted under the fluids conservative regime align with expected outcomes.

Readers should exercise caution when interpreting these plots in Figure 2 as indicators of treatment effects. Simulated trajectories might terminate prematurely within the 24-hour period due to events such as death or discharge. Consequently, the number of patients (or trajectories) contributing to the average may not remain constant throughout the entire 24-hour period.

Individual Patient Counterfactual Prediction In this section, we present the results of applying the G-Net LSTM model, the overall superior performer during our observational checks, to the task of counterfactual predictions for an individual patient from our test set. The patient is a 75 year old female with a history of congestive heart failure (CHF), admitted to MICU/SICU with 4L of pre-ICU fluids. Within 4 hours after ICU admission, patient had developed a hypotensive episode (with mean arterial blood pressure $<$ 60mmHg), and received an additional fluids bolus (1L) in the ICU. The counterfactual strategies were applied starting at the end of hour 5 after ICU admission. In Figure 3, we highlight the contrast in predicted counterfactual outcomes under a fluid liberal regime (green) compared to a fluid conservative regime (blue), which imposes a cap of 5L in total volume (pre-ICU and in-ICU) in fluid bolus administration. Additionally, we show the predicted trajectory of the patient under the observational regime (magenta), where fluid boluses were administered based on G-Net's prediction conditioned on the patient's history. We perform 100 Monte Carlo simulations under each fluid resuscitation strategy. Figure 3 shows that there was more bolus administered to the patient under the fluid liberal strategy (green) compared to fluid conservative strategy (blue). Under

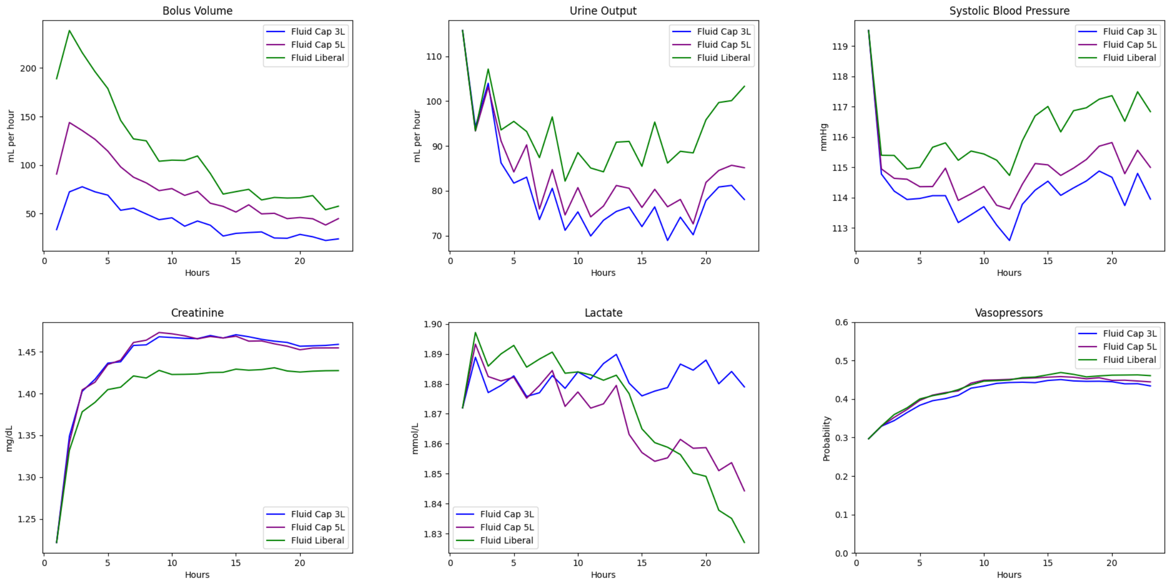

Figure 2: Population-level trajectories under counterfactual fluids strategies. G-Net prediction for selected covariates under various counterfactual fluids strategies, including two fluids limiting strategies with 3L (blue) and 5L (purple) fluids cap, and a fluids liberal (green) strategy without fluid cap.

the fluid liberal strategy, the fluid balance of the patient progressively increases over time and diverges further from that of the conservative fluid regime. This is accompanied by a rising probability of the patient requiring mechanical ventilation, indicative of a potential fluid overload compared to the more conservative approach.

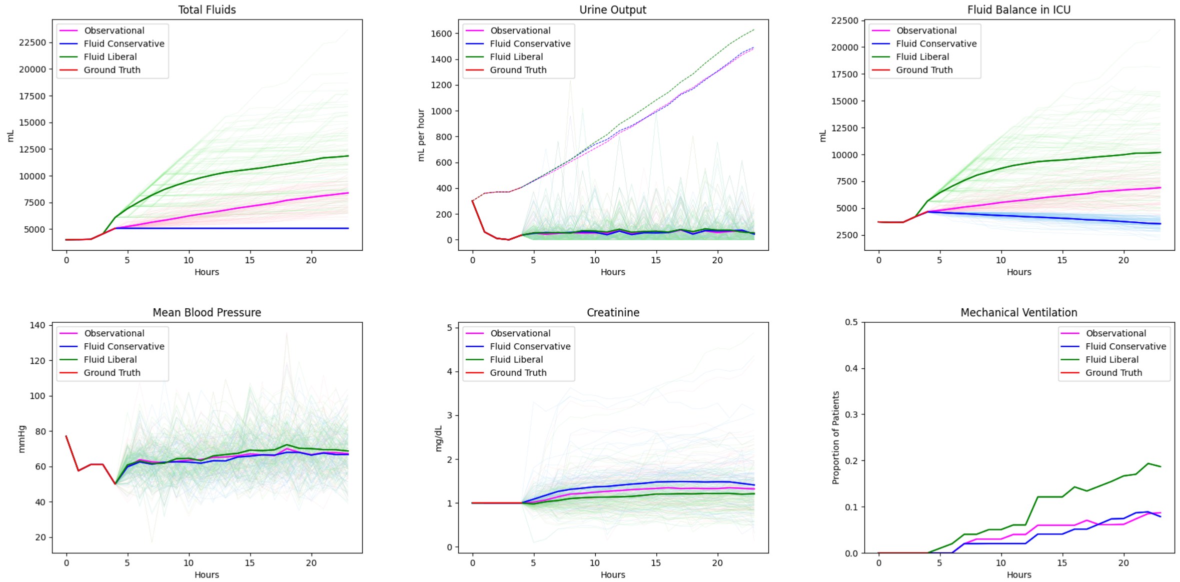

Figure 3: MIMIC case study: Counterfactual prediction for an example sepsis patient under observational (magenta), fluid liberal (green) vs. fluid conservative with 5L fluid cap (blue) counterfactual fluid regimes. Each plot shows 100 Monte Carlo simulations in light colors, average in solid dark colors for the patient under the observational and two different CF strategies (starting t = 5). Ground-truth observations from the initial 5 hours in the ICU in red. Predicted trajectory under observational regime in magenta. Urine output also show cumulative values in dashed lines.

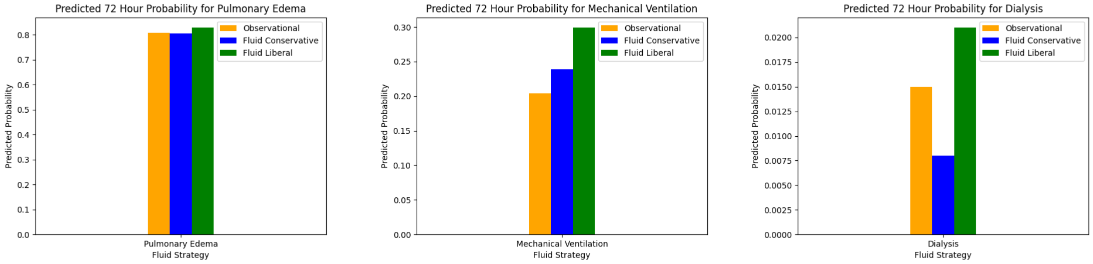

Figure 4: MIMIC case study: Predicted probability of patient experiencing adverse outcomes associated with fluid overload within 72 hours of ICU admission under the observational, fluid liberal, and fluid conservative regimes.

Figure 4 displays the predicted probabilities of this patient developing various adverse clinical outcomes under the observational, conservative and liberal fluid strategies respectively. We note that the patient is predicted to have a higher probability of experiencing each of the adverse outcomes of interest associated with fluid overload under the fluid liberal strategy compared to the conservative strategy. Nevertheless, one should use caution when interpreting these results as measures of treatment effects, given the absence of uncertainty quantification. Instead, this example serves to illustrate a potential clinical use case for our approach.

5 Discussion & Conclusion

In this paper, we present an application of G-Net, a deep learning implementation of g-computation, to estimate clinical outcomes under counterfactual treatment strategies in a real-world cohort of sepsis patients. We assessed the predictive abilities of G-Net on a hold-out test set under the observational regime and demonstrated that the model is able to accurately simulate forward most covariates at testing time. When tested under varying length of treatment history, G-Net maintains robust performance across our experimental settings.

When we applied G-Net to analyzing outcomes under various counterfactual treatment strategies, we found that G-Net made counterfactual predictions that are aligned with clinical expectations. One limitation of this work is that, counterfactual predictive density estimates in our experiments do not take into account uncertainty about model parameter estimate. Specifically, given G-Net parameters, the distribution of the Monte Carlo simulations produced by G-Net constitute an estimate of uncertainty about a counterfactual prediction. However, this estimate ignores uncertainty about the G-Net parameter estimates themselves. One way to incorporate such model uncertainty would be to fit a Bayesian model and, before each Monte Carlo trajectory simulation in g-computation, draw new network parameters from their posterior distribution. These Monte Carlo draws would be from the posterior predictive distribution of the counterfactual outcome. Bayesian deep learning can be prohibitively computationally intensive, but can be approximated through dropout [20]. An important area of future work for G-Net is adding support for quantification of model uncertainty.

The findings presented in this study suggest promising potential for G-Net's clinical utility in enhancing treatment decision-making for sepsis patients in the ICU. While our primary focus centered on clinical applications, we acknowledge that G-Net can be adapted for a variety of scenarios in which one would like to predict downstream effects of alternative courses of action.

6 Acknowledgements

The authors thank Professors Zach Shahn and Roger Mark for their insightful comments, and Amy Hu for her assistance. Lehman is in part funded by MIT-IBM Watson AI Lab, NIH grants R01EB030362 and 1U01TR003528-01A1.

References

1. Hernan MA, Robins JM. Causal inference: What if. 1st ed. Chapman & Hall CRC; 2011.

2. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure perio-dapplication to control of the healthy worker survivor effect. Mathematical modelling. 1986;7(9-12):1393–1512.

3. Li R, Hu S, Lu M, Utsumi Y, Chakraborty P, Sow D, et al. G-Net: a Recurrent Network Approach to G-Computation for Counterfactual Prediction Under a Dynamic Treatment Regime. In: Proceedings of Machine Learning for Health. vol. 158; 2021. p. 280-97.

4. Johnson A, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Nature Scientific Data. 2023.

5. Taubman S, Robins J, Mittleman M, Hernan M. Intervening on Risk Factors for Coronary Heart Disease: An Application of the Parametric G-Formula. International journal of epidemiology. 2009.

6. Young J, Cain L, Robins J, O'Reilly E, Hernan M. Comparative Effectiveness of Dynamic Treatment Regimes: An Application of the Parametric G-formula. Statistics in biosciences. 2011.

7. Shahn Z, Shapiro NI, Tyler PD, Talmor D, Lehman LH. Fluid-limiting treatment strategies among sepsis patients in the ICU: a retrospective causal analysis. Journal of Critical Care. 2020;24(62).

8. Early Restrictive or Liberal Fluid Management for Sepsis-Induced Hypotension. New England Journal of Medicine. 2023;388(6):499-510. PMID: 36688507. Available from: https://doi.org/10.1056/NEJMoa2212663.

9. Lim B, Alaa A, van der Schaar M. Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018. Available from: https://proceedings.neurips.cc/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.

10. Alaa MA, Weisz M, van der Schaar M. Deep Counterfactual Networks with Propensity-Dropout. In: Proceedings of the 34th International Conference on Machine Learning (ICML); 2017. .

11. Atan O, Jordan J, Van der Schaar M. Deep-Treat: Learning Optimal Personalized Treatments from Observational Data using Neural Networks. In: Proceedings of AAAI; 2018. .

12. Yoon J, Jordan J, Van der Schaar M. GANITE: Estimation of Individualized Treatment Effects Using Generative Adversarial Nets. In: ICLR; 2018. Available from: https://openreview.net/forum?id=ByKWUeWA-.

13. Bica I, Alaa AM, Jordon J, van der Schaar M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. International Conference on Learning Representations. 2020.

14. Schulam P, Saria S. Reliable Decision Support using Counterfactual Models. In: Neural Information Processing Systems (NIPS).; 2017. .

15. Xu Y, Xu Y, Saria S. A Bayesian Nonparametric Approach for Estimating Individualized Treatment-Response Curves. In: Proceedings of the 1st Machine Learning for Healthcare Conference (PLMR). vol. 56; 2016. p. 282-300.

16. Melnychuk V, Frauen D, Feuerriegel S. Causal Transformer for Estimating Counterfactual Outcomes. In: Proceedings of the 39 th International Conference on Machine Learning; 2022. .

17. Wu F, Zhao G, Zhou Y, Qian X, Baedorf-Kassis E, Lehman LH. Forecasting Treatment Outcomes Over Time Using Alternating Deep Sequential Models. IEEE Transactions on Biomedical Engineering. 2023.

18. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016 Feb;315(8):801–810.

19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45.

20. Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks. In: Advances in Neural Information Processing Systems; 2016. .